Quarterly newsletter

CHAIR

iDANAE

INTELLIGENCE · DATA · ANALYSIS · STRATEGY

3Q20

MLOps, a key element in the digital ecosystem

POLITÉCNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

ManagementSolutions
Making things happen

Analysis of meta-trends

The iDanae Chair for Big Data and Analytics (where iDanae stands for intelligence, data, analysis and strategy in Spanish), created within the framework of a collaboration between Universidad Politécnica de Madrid (UPM) and Management Solutions, aims to promote the generation and dissemination of knowledge, the transfer of technology, and the furthering of R&D in the Analytics field.

One of the lines of work developed by the iDanae Chair is the analysis of meta-trends in the field of Analytics. A meta-trend can be defined as a value-generating concept or area of interest within a particular field that will require investment and development from governments, companies and society in the near future[1].

This report is focused on the evolution of data science ecosystem and the concept and relevance of Machine Learning operations, or MLOps.

[1]Para más información, véase la primera publicación de iDanae (iDanae, 3Q 2019).

Introduction: creating a data driven culture

Due to the increase in computer power and the development of new advanced analytical tools that facilitate the processing of large amounts of data, an increasing number of companies have specific teams specialised in data analysis that make use of machine learning techniques. Ideally, these teams should be integrated with the rest of the business units of the companies, generating a trend towards the democratisation of data[2]. The ultimate objective is that a large number of users within the company can enrich their decision-making process with the information extracted from the available data, without the need of having great technical knowledge or direct dependence on a specialised profile.

A new way of collaboration is emerging among the different trends related to the democratisation of data. This practice promotes the creation of an ecosystem with the aim of facilitating and standardising cooperation between data science teams in different areas. Among others, some examples of these areas include knowledge generation and dissemination (via platforms such as Medium or Towards Data Science), resolution of incidents or questions (through platforms such as Stack Overflow), the development of libraries (such as XGBoost), code and projects (through the use of tools such as Jupyter Notebooks or Github), or the resolution of specific problems (for example, in platforms such as Kaggle or the conduction of Hackathons).

This ecosystem has been favoured by the sectoral specialisation, so that new roles have emerged to offer increasingly elaborate solutions: currently, a distinction is made between profiles such as data analyst, data scientist, data engineer, or data architect, among others. Despite this, the ecosystem is still under development, and has yet to adapt to the demand, and offer solutions to the different problems that arise when integrating it into a company's business, such as, for example, those that arise around the extraction of information from data or during the modelling process. This is accompanied by various related debates, such as the ethics in the development of artificial intelligence projects, or the limitations of these systems because of the mathematical complexity of the models used or the quality of the training data[3].

One of the practices that is becoming increasingly relevant is Machine Learning Operations or MLOps, which arises from collaboration between data science and IT teams. MLOps is defined as the set of practices and tools used for the development and validation of machine learning models and their efficient and smooth implementation[4]. These tools cover the areas of programming and organising the model development and implementation function, and can be collaborative. Although there could be an analogy in the emergence of DevOps[5] (a set of best practices and tools used for the development, improvement and testing of applications, software or services, and their effective implementation, at high speed) in the late 1990s, the concept of MLOps incorporates a set of additional challenges that need to be considered. In the case of DevOps, the aim is to increase the efficiency in the production of the developed systems, speeding up the production process so that the developer does not have to worry about the effective implementation. DevOps is then oriented towards the integration of the software into the operations of the IT areas, which may not be aligned with the business operations. The appearance of MLOps adds a level of complexity by incorporating the need of ensuring data quality across all areas of the company. Therefore, data quality and governance management must be carried out transversally to business operations, and this must be aligned with efficiency in the implementation process, which is more closely linked to IT operations.

This practice arises in response to the wide variety of machine learning solutions in the industry, as well as the need of (i) ensuring efficient collaboration between development teams, (ii) producing a complete and updated set of documentation, (iii) achieving version control, (iv) reducing implementation time, and (v) carrying out adequate monitoring once the developed model has been set into production, among others. In fact, it is expected that by 2025 the MLOps market will exceed 4 billion dollars[6], and is already a requirement for inclusion in Gartner's Magic Quadrant[7].

[2]This trend has been the subject of study in a previous publication, please refer iDanae, 1Q 2020.
[3]Both trends have been covered in previous publications, please refer iDanae, 2Q 2020 and iDanae, 4Q 2019, respectively.
[4]Talagala, 2018. This concept first appeared in an article published by Google in 2015 (Sculley, D. and others, 2015), although the term began to be coined from 2018 onwards following a speech given by Google itself (Kaz Sato, 2018).
[5]DevOps is an acronym for the words development and operations. It arises with the aim of unifying software development (Dev) and software operation (Ops). The term became popular in 2009, with the DevOps Days events initially held in Ghent, Belgium (DevOps Days, 2009).
[6]Cognilytica, 2020.
[7]Gartner, 2020.

MLOps: an integrated working approach

Currently, the processes run by both data science and operations teams are separated in many organisations. Moreover, the integration of the models into the systems and their production is done in the last phase of the process (figure 1). This approach generates risks and possible errors that can affect the production process of the models developed, which can lead to the failure of the model implementation. For example, currently only 14% of projects are implemented in a period shorter than seven days, while this process takes over 25% of the time of data science teams[8]. In addition, modifications, adjustments or retraining are generally needed once a model has been implemented due to changes in the quality or type of data used, in the application portfolio, in the conditions of use, or in the business strategy itself. These modifications need to be made effectively and quickly[9].

The goal of the MLOps teams is to establish best practices that enable a model to be replicable, collaborative, scalable and automated[10]. To do this, they seek to integrate the development of the models with the company's operations process (figure 2), so that there is greater collaboration and integration between the operations and data science teams[11]. MLOps, therefore, considers the entire life cycle of the model jointly, paying special attention to the deployment, as this is the phase in which there should be the greatest interaction between both teams with the aim of improving the quality and coherence of the adopted solutions. To this end, on the one hand, it is essential to monitor and test all the processes, the data used, the execution of the models and the start of production by means of continuous integration and distribution systems (CI/CD); on the other hand, a statistical follow-up must be established to warn of the situations in which it is necessary to modify the model (for example, when it loses predictive capacity, or when the input data undergoes some modification). There must also be model governance, which facilitates the audit, compliance, access and security functions[12].

[8]Algorithmia, 2020.
[9]INNOQ, 2020).
[10]Kobran, 2020
[11]Open Data Science, 2020.
[12]Forbes, 2020.
[13]McKnight, 2020.
[14]McKnight, 2020.

Figure 1: The usual process of developing and putting into production machine learning models in many organisations is separate, generating risks, delays and errors[13].

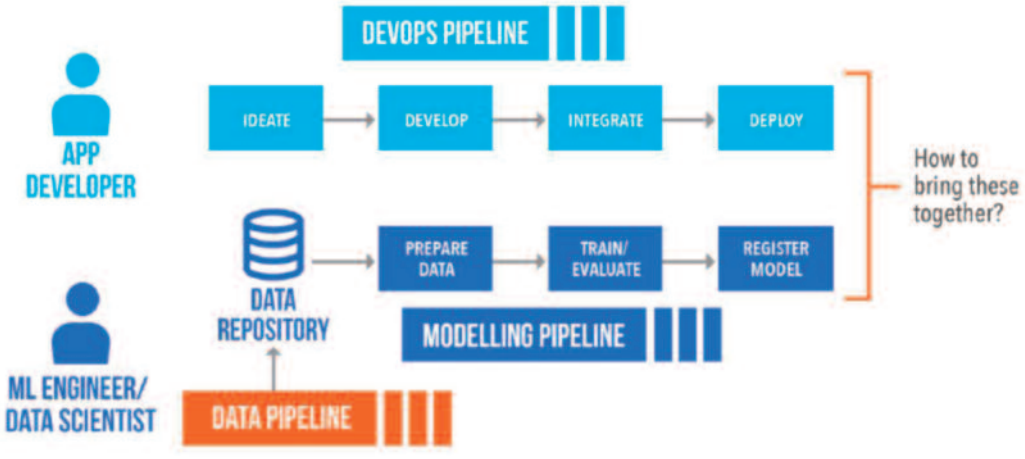
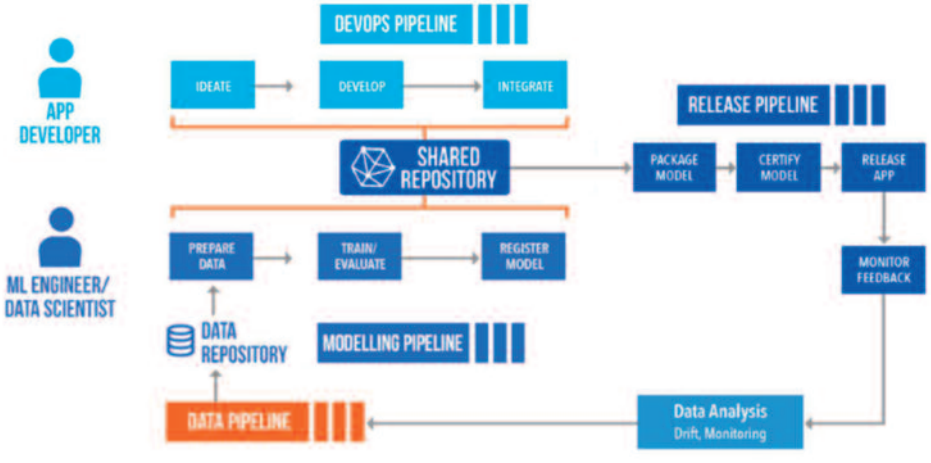


Figure 2: MLOps proposes to integrate the data science teams with the operations teams, generating a unique collaborative flow, allowing the creation of reproducible, scalable and more automated models[14].



MLOps drivers

MLOps is characterised by four fundamental pillars:

- reproducibility, with the use of measures such as object oriented programming, encapsulation of objects in APIs, or the use of containers;
- scalability, through the use of specific infrastructures, such as cloud infrastructures, with self-scalable and parallelisable systems;
- the creation of collaborative spaces, through the use of services such as Git or tools for managing data and the life of the model; and
- continuous integration through the automation of testing and the deployment of applications.

A set of best practices is established around these four pillars to ensure the quality and production of the models developed. These practices are based on the four main components of the model life cycle[15]: data management, model development, infrastructure and model validation and monitoring. Seven tests are established on each of these components to ensure the correct functioning of all the processes.

Data management

Machine learning models behave differently depending on the data used for training. For this reason, it is necessary to have exhaustive control over the data sources, their quality and their evolution over time. The following elements have to be analysed:

1. The variables behave as expected: the variables must show reasonable statistical values, as well as being within expectations if there is expert knowledge of the variables used.

2. All the variables used improve the result: it is not recommended to add all the available variables, but only those that improve the predictive power (which generally means eliminating those that are highly correlated with other variables used).

3. No variable is expensive: it is not recommended to use variables that are difficult to obtain or to treat at a computer level in relation to the improvement of the final predictive power of the model.

4. Variables comply with predefined policies: prerequisites on the type of information used (e.g., race, religion, sexual orientation, etc.) must be established. This information should not be inferred from other variables.

5. Data management has adequate privacy controls: despite the fact that data is anonymised, the transformation and processing of variables can lead to possible identification, and it is necessary to establish controls to detect and avoid this.

6. New variables can be quickly incorporated: the faster new ideas can be incorporated into the model in production, the sooner the benefits can be realised.

7. All the code is tested: it is necessary to establish tests on the codes used for obtaining and createing variables, in order to find errors and incorrect operations, avoid unwanted inferences and reduce computational costs.

Model development

As it is done for software development, best practices must also be established for the development of machine learning models, so that code executions and updates are controlled, all configuration parameters are monitored, and the degradation that will occur over time is acknowledged:

1. All model specifications are reviewed and stored in a repository: it is not recommended to make runs using the modifications made by a particular user, but they should be reviewed first by other team members to ensure their quality. In addition, it is necessary to know which exact version of the model has been used in each run in order to be able to reverse it in case the results get worse, and be able to learn from all the runs.

2. Proxies used offline must be correlated with online metrics: it is advisable to know the relationship between typical metrics used in machine learning models (such as quadratic error) and impact metrics to achieve a model with better results.

[15] Breck, Cai, Nielsen, Salib, & Sculley, 2017.

3. All hyperparameters have been adjusted: there are several search strategies such as grid search or Bayesian search that allow the adjustment of hyperparameters, so that the model works in an optimal way.

4. The impact of model obsolescence is controlled: models in production must be trained quickly enough to ensure that they are kept up to date, which is particularly important in the presence of non-stationary data. It is necessary to know how the predictive capacity of the model degrades in cases where it is not possible to update the models so quickly, establishing thresholds in the predictive capacity, from which the model is updated.

5. The simplest model is being used: it is recommended to use the simplest techniques that provide sufficient predictive capacity, which allows for cost savings, improved interpretability and, in certain areas, coverage of regulatory requirements.

6. The quality of the model on relevant sections or subpopulations of data is as required: there may be global metrics that hide problems at a higher level of granularity. The use of relevant data sections is recommended.

7. The model is inclusive: due to problems with training data, biases and discriminations on certain population groups may arise.

Infrastructure

Machine learning models usually require a complex infrastructure, so it is advisable to carry out tests to ensure the proper functioning of the platforms used, so that the model works in the same way in the development environment as in the production environment. The following elements must be assessed:

1. Training is replicable: if a model is trained several times on the same dataset, the same resulting model should be obtained (deterministic training). Although this is not always possible, it should be attempted in all situations where it is, and reduce indetermination.

2. The model code is tested: in the same way as with the code for obtaining and creating variables, all model code must be tested for possible bugs generated in its development.

3. There are tests throughout the integration process: each phase of the process can introduce errors that affect the following phases, so each phase need to be tested.

4. The quality of the model is validated before it is implemented into production: once the model has been trained, and before it is implemented in the production environment, it is necessary to ensure the quality of the model in order to accept or reject this implementation.

5. It is possible to run the model code for debugging: there must be a simple and well-documented process for understanding possible unexpected behaviour, and the output of the model should be controlled under unexpected data inputs.

6. Deployment of new models simultaneously with the old ones: It is recommended that the implementation of a new model is done simultaneously with the previous version, receiving a small portion of the total data, which allows to check the correct functioning before completely replacing the previous version.

7. Updating a model can be reversed: if a model starts working unexpectedly, it should be possible to use the previous version again in a simple, quick and efficient way.

Validation and monitoring

It is essential that the model implemented works properly over time. To do this, it is necessary to be able to enlarge the training dataset with new data, as well as to control both the data on which predictions are executed, and the predictive capacity of the model. The following elements must be verified:

1. Reporting when input data are modified: updates or changes in the source of input data may distort variables or change their meaning, which may imply the need to re-train the model or develop new functionalities.

2. Training data have the same properties as prediction data: different sets of input data need to be analysed to ensure comparability, since prediction data, being more recent, may have undergone changes with respect to training data.

3. The variables created must be the same during training and during prediction: it is necessary to avoid that the variables obtained through feature engineering processes are different during training and during prediction, or that they are modified when introducing new variables.

4. The models are not obsolete: reference metrics must be available and the passage of time and its impact on the model's predictive capacity must be monitored.

5. The models are numerically stable: the models do not offer out-of-range, null or infinite values during training or during the prediction phase.

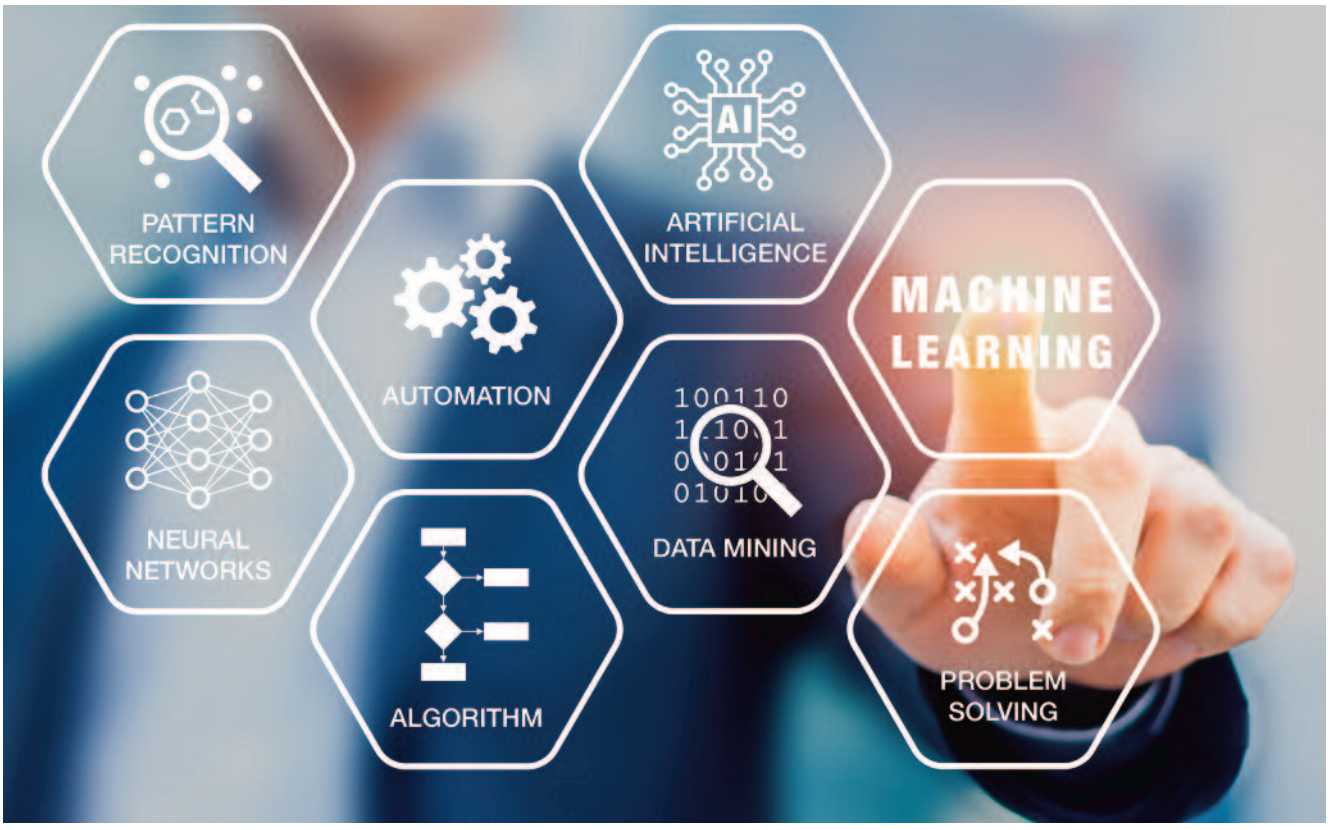
6. The computational cost remains constant: the model should not take longer to run or require more resources than during its development, or throughout its life cycle.

7. The quality of the prediction remains constant: the validation data are usually older than the real data, and therefore the measured quality of the model is only an estimate of the real quality. It must be ensured that, once put into production, the results of the model remain of the expected quality.

Conclusions

MLOps offers an integrated working approach, with the teams involved in the development and implementation of machine learning models playing a fundamental role in companies' adoption of data driven culture, ensuring a more deterministic, automated and scalable process. This will generate benefits in the modelling processes, such as a higher success rate when developing machine learning projects, greater reproducibility of the models (which also implies greater ease in interpreting and explaining the model's predictions) and greater speed and agility in converting new ideas into models effectively implemented. However, this requires a transformation to address data quality in a centralised way, mainly when data collection, storage and maintenance are not centralised and ML processes require integration of data from different business operations.

Analogous to DevOps, MLOps relies on comprehensive testing and monitoring of all elements related to the development of the model and its production, including data used, modelling, infrastructure, or model degradation over time, encouraging the automation of this process to the greatest extent possible. This allows a standardisation and quality improvement of the model creation process, resulting in competitive advantages for companies.



Bibliography

Algorithmia. (2020). 2020 state of enterprise.

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction. Proceedings of IEEE Big Data.

Cognilytica. (2020). ML Model Management and Operations 2020 ("MLOps").

DevOps Days. (2009). DevOps Days. Retrieved from DevOps Days: https://devopsdays.org/

Forbes. (2020). The emergence of MLOps. Obtained from Forbes: https://www.forbes.com/sites/cognitiveworld/2020/03/08/the-emergence-of-ml-ops

Gartner. (2020). Magic Quadrant for data science and machine learning platforms.

iDanae. (1Q 2020). Data democratization.

iDanae. (2Q 2020). Limits of modelling.

iDanae. (3Q 2019). Interpretability of artificial intelligence models.

iDanae. (4Q 2019). Ethics and artificial intelligence.

INNOQ. (2020). Why you Might Want to use Machine Learning. Obtained from Machine Learning Operations: https://ml-ops.org/content/motivation.html

Kaz Sato. 2018. Speech entitled "What is ML Ops? Best Practices for Devops for ML".

Kobran, D. (2020). What is MLOps? Obtained from Paperspace: https://blog.paperspace.com/what-is-mlops/

Management Solutions. (2020). Auto Machine Learning, towards model auomatization.

McKnight, W. (2020). Delivering on the Vision of MLOps.

Open Data Science. (2020). What are MLOps and Why Does it Matter? obtained from Medium: https://medium.com/@ODSC/what-are-mlops-and-why-does-it-matter-8cff060d4067

Sculley, D.; Holt, Gary; Golovin, Daniel; Davydov, Eugene; Phillips, Todd; Ebner, Dietmar; Chaudhary, Vinay; Young, Michael; Crespo, Jean-Francois; Dennison, Dan. 2015. "Hidden Technical Debt in Machine Learning Systems", NIPS Proceedings.

Talagala, N. (30 de January de 2018). Why MLOps (and not just ML) is your Business' New Competitive Frontier. AITrends.

Authors

Ernestina Menasalvas (UPM)

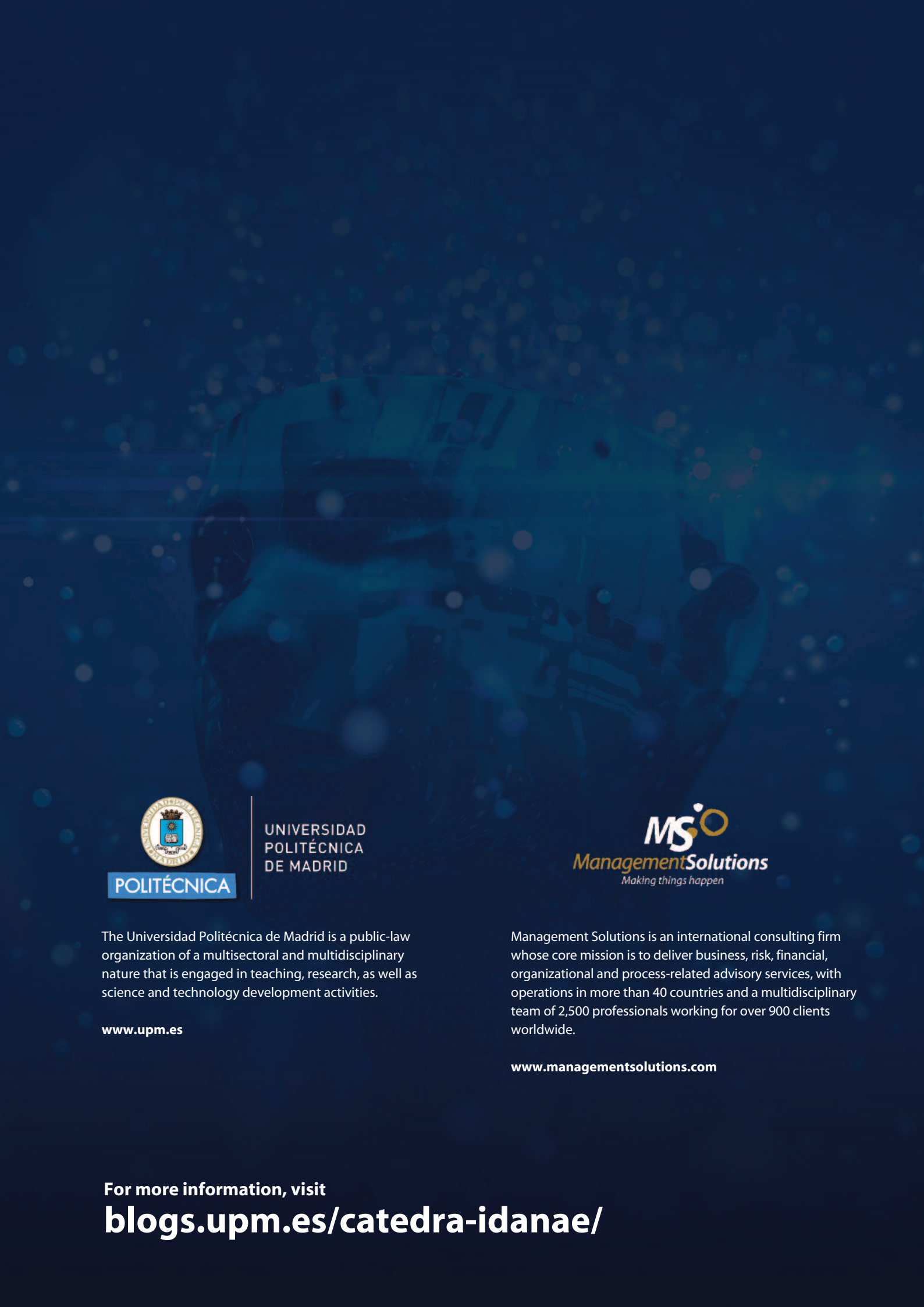
Alejandro Rodríguez (UPM)

Manuel Ángel Guzmán (Management Solutions)

Daniel Ramos (Management Solutions)

Segismundo Jiménez (Management Solutions)

Carlos Alonso (Management Solutions)



UNIVERSIDAD POLITÉCNICA DE MADRID

The Universidad Politécnica de Madrid is a public-law organization of a multisectoral and multidisciplinary nature that is engaged in teaching, research, as well as science and technology development activities.

www.upm.es

Management Solutions is an international consulting firm whose core mission is to deliver business, risk, financial, organizational and process-related advisory services, with operations in more than 40 countries and a multidisciplinary team of 2,500 professionals working for over 900 clients worldwide.

www.managementsolutions.com

For more information, visit

blogs.upm.es/catedra-idanae/